DOCUMENT RESUME

ED 378 195 TM 022 525

AUTHOR Tang, Huixing
TITLE A Simultaneous Approach to Multi-Factor DIF Analysis.
PUB DATE Apr 94
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 5-7, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Analysis of Variance; *Difficulty Level; Error of Measurement; *Item Bias; Item Response Theory; Simulation; *Test Items
IDENTIFIERS *Multifactorial Models; *Residual Scores

ABSTRACT

A method is presented for the simultaneous analysis of differential item functioning (DIF) in multi-factor situations. The method is unique in that it combines item response theory (IRT) and analysis of variance (ANOVA), takes a simultaneous approach to multifactor DIF analysis, and is capable of capturing interaction and controlling for possible confounding variables. It is referred to as the IRT-ANOVA method. The most salient feature is that the procedure used IRT to control for group ability differences and familiar inferential procedures to test the DIF effect. Residuals can be construed as item scores free from the effects of both personal ability and item difficulty. The use of ANOVA provides not only a test statistic based on a familiar distribution, but also descriptive measures of DIF magnitude in terms of group means and variances. Simulations in the one-factor, two-group situation reveal the usefulness of the approach and indicate error rates. Seven figures illustrate the simulations. (Contains 8 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
HUIXING TANG
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 378 195

A Simultaneous Approach to Multi-Factor DIF Analysis

by

Huixing Tang

The Psychological Corporation

Paper presented at the 1994 annual meeting of the
National Council on Measurement in Education
New Orleans, Louisiana

TM022525

BEST COPY AVAILABLE

I. RATIONALE

Popular methods for analysis of differential item functioning (DIF), by definition and design as well as in practice, are confined to the investigation of one DIF factor (Hills, 1989; Cole & Moss, 1989). Even with methods capable of dealing with 'non-uniform' DIF, attention is focused on only one DIF variable in relation to different regions of the ability continuum. When two or more factors such as gender, ethnicity, and social economic status need to be investigated, they are typically analyzed separately, even though the factors under investigation are known to be related to, or interact with, each other. It is probably not inappropriate to characterize popular DIF methods as belonging to "the pre-factorial era," when "the law of single factor" prevailed (Fisher, 1937).

The limitations of the "one-factor" approach are well documented in the literature on experimental design. With respect to DIF analysis, the one-factor approach is not only incapable of capturing interaction effects, but may also lead to misleading results regarding main-effect DIF. An item may be flagged as biased when the detected effect is, in fact, a function of a confounding variable. On the other hand, a biased item may fail to be flagged when the effect is cancelled out or its magnitude reduced by an intervening variable not included in the design.

Educational and psychological tests typically measure constructs related to a multitude of factors which are known to interact with each other to varying degrees. DIF analysis, embedded in such a context, ought to be based on sound theorization of the causative factors of the behavior being measured and the relationship between these factors. Methodologically, efforts should be made to develop procedures that are capable of dealing with multiple factors and are, at the same time, sensitive to their complex interrelationships.

The purpose of this paper is to present a simultaneous method for DIF analysis in multi-factor situations. The method is unique to the existing methods in that it combines item response theory and analysis of variance, takes a simultaneous approach to multi-factor DIF analysis, and is capable of capturing interaction and controlling for possible confounding variable(s). As the method employs both IRT and ANOVA, it will be referred to as the IRT-ANOVA method.

II. THE PROCEDURE

The IRT-ANOVA method consists of the following steps:

1. Calibrate the data with an appropriate IRT model to obtain estimates of person ability and item difficulty, or step difficulty for polychotomously scored items.

2. Using the estimates obtained in step 1, compute P_{ij}, the probability of person i responding correctly to item j, or P_{ijk}, the probability of person i scoring in the kth category of item j.

3. Using the probabilities obtained in step 2, compute E_{ij}, the expected score for person i on item j. For dichotomous items, E_{ij} is equivalent to P_{ij}. For polychotomous items with m+1 categories, E_{ij} is computed by

$$E_{ij}=\sum_{k=0}^{m_j} kP_{ijk}$$

4. Compute R_{ij}, the residual score for person i on item j by subtracting E_{ij} from the observed item score X_{ij}

$$R_{ij}=X_{ij}-E_{ij}$$

or Z_{ij}, the standardized residual score for person i on item j by dividing R_{ij} by the standard deviation of X_{ij}

$$Z_{ij}=\frac{R_{ij}}{\sqrt{\sum_{k=0}^{m_j}(k-E_{ij})\,P_{ijk}}}$$

For dichotomous items, the equation can be simplified to

$$Z_{ij}=\frac{R_{ij}}{\sqrt{P_{ij}(1-P_{ij})}}$$

5. Perform analysis of variance on the data with R_{ij}'s or Z_{ij}'s as values of the dependent variable and DIF factor(s) under investigation as the independent variable(s), and use the resulting F ratio as the test statistic for DIF.

6. Compute the marginal or cell mean of the residuals for each group as well as the differences between the means as measures of observed effect size.

The most salient feature of the procedure is that it uses IRT to control for group ability differences and familiar inferential procedures to test the DIF effect. Residuals can be construed as item scores free from the effects of both person ability and item difficulty. They are expected to be random with a mean of 0. A positive residual may imply that the person is scoring higher than expected based on overall test performance. A negative residual may imply that the person is scoring lower than expected. Consistently high (or low) residual values for a particular subgroup may imply that the item favors (or disfavors) the examinees in the subgroup. The use of ANOVA provides not only a test statistic based on a familiar distribution, but also descriptive measures of DIF magnitude in terms of group means and variances.

The IRT-ANOVA has many desirable features. It is capable of:

- simultaneously processing multiple factors under investigation
- simultaneously processing multiple levels of a studied factor
- simultaneously processing dichotomous and polytomous items
- examining interaction effects and controlling for confounding
- controlling for group ability differences at each examinee level
- providing a test statistic using a familiar inferential procedure
- providing easily interpretable descriptive measures of DIF
- being simple and straightforward for use and understanding
- being easily replicable for research or methodological inquiry
- allowing for relatively small sample sizes for the focal group
- avoiding scale shift due to separate calibration for each group
- avoiding loss of within-group information by ability grouping

Tang (1994) investigated the use of the IRT-ANOVA in the one-factor, two-group situation. The relevant findings of the simulation studies include:

1. In small sample situations, the IRT-ANOVA method is more powerful than the Mantel-Haenszel method when the data fit the model or when the data do not fit the model but the impact of misfit resulting from lack of unidimensionality is randomly distributed between the two groups.

2. Model-data misfit as a function of dimensionality has little or no effect on power when the impact of the secondary dimension is evenly or randomly distributed between the groups. When the effect of the secondary dimension is differentially distributed between the groups, however, power increases if the secondary dimension favors the group the item favors and decreases if the

secondary dimension favors the group the item disfavors.

3. The error rate is close to its nominal level when there is no group ability difference, whether or not the data fit the model.

4. As the sample size increases, there is a slight monotonic increase in error rate when there is a group ability difference. There is a considerable increase in error rate (from 5 to 30 percent) when the second dimension has differential impact on the two groups.

III. DESIGN OF THE SIMULATION STUDIES

A simulation study was conducted to investigate the use of the IRT-ANOVA method for DIF analysis involving two factors. The main purpose of this study was to examine the effect of sample size, group ability difference, and model-data misfit on the power and the error rate of the method in detecting DIF with interaction effects.

The examinee samples were simulated using the unit normal distribution. The sample sizes varied from 200 to 1400 in increments of 200. The DIF factors simulated were gender (Male and Female) and ethnicity (White and Black). The group sizes were equal for both factors. Item difficulties for 40 dichotomously scored items were simulated using the unit normal distribution. They were randomly generated for each replication so that the effect of item difficulty on DIF was controlled. Six DIF items were introduced, with two items for each of the following three types:

1. *main-effect*: the marginal mean of one level is higher than the marginal mean of the other level
2. *ordinal interaction*: the cell means associated with the levels of one factor occupy the same ordinal position at each level of the other factor, but differ in magnitude
3. *disordinal interaction*: the cell means associated with the levels of one factor do not occupy the same relative positions over levels of the other factor (Kennedy & Bush, 1978).

The main-effect DIF was introduced in items 1 and 2, with item 1 favoring (being easier for) females and item 2 favoring males. The ordinal interaction DIF was introduced in items 3 and 4, with item 3 favoring white females and item 4 favoring black females. The disordinal interaction DIF was introduced in items 5

and 6, with item 5 favoring white females and black males and item 6 favoring black females and white males. The DIF magnitude was set at .6 in logit difficulty for all DIF items (see Figure 1 for a graphic display of the DIF items).

The simulated DIF items were examined under the following conditions:

1. equal group ability and data fit the model
2. unequal group ability and data fit the model
3. equal group ability and data do not fit the model
4. unequal group ability and data do not fit the model

Group ability difference was simulated for males with a mean logit ability .6 higher than the mean logit ability of females. Model-data misfit was introduced by simulating the responses to all the DIF items and one non-DIF item using a secondary ability distribution (B_2) correlated imperfectly (r=.5) with the primary ability distribution (B_1) that generated the responses for the rest of the items. B_1 and B_2 were generated as follows(Hogg & Craig, 1978, p. 143):

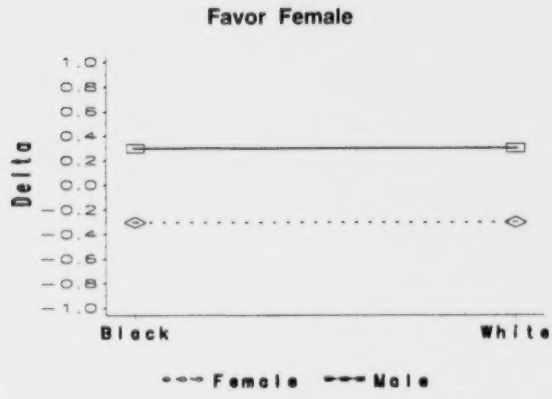
$$B_1 = X_1$$
$$B_2 = RX_1 + (1 - R^2)^{1/2}X_2$$

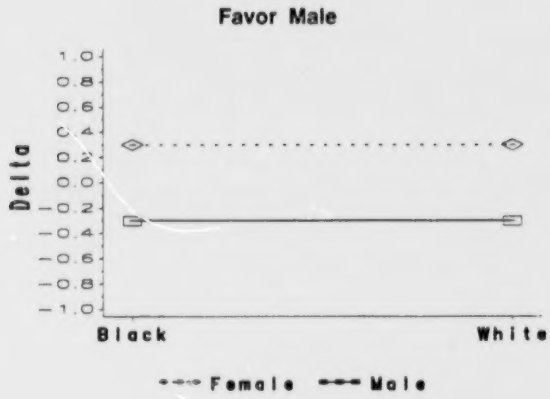
where X_1 and X_2 are independent random variables with a standard normal distribution and R is the correlation between B_1 and B_2.

It should be noted that under conditions where there is a group ability difference, the secondary distribution was simulated after the group ability difference was introduced. Consequently, the secondary ability distribution invariably favors the low ability group (females) and disfavors the high ability group (males), due to the regression-toward-the-mean effect caused by the imperfect correlation between the two distributions. Where there is no group ability difference, however, the impact of the secondary dimension is expected to be distributed evenly among the gender groups. The misfit thusly simulated allows for examination of its relationship with DIF detection when the misfit *does* and *does not* have a differential impact on the groups being compared.

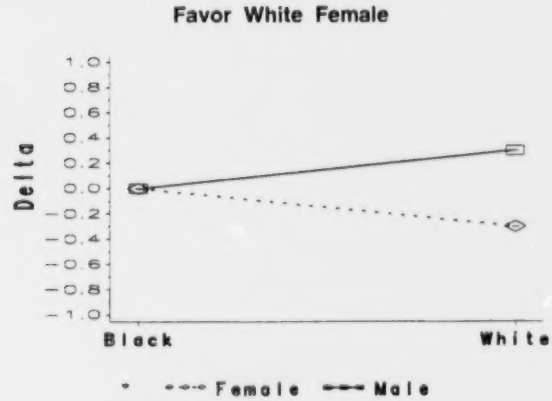
Figure 2 presents a graphic display of all possible combinations of *item*, *sample size*, and *condition*. Each cell was replicated 100 times. The Rasch model was used for both data simulation and IRT calibration. A computer program was written by the author which processes data simulation, IRT calibration using the unconditional maximum likelihood method, and computation of ANOVA statistics.

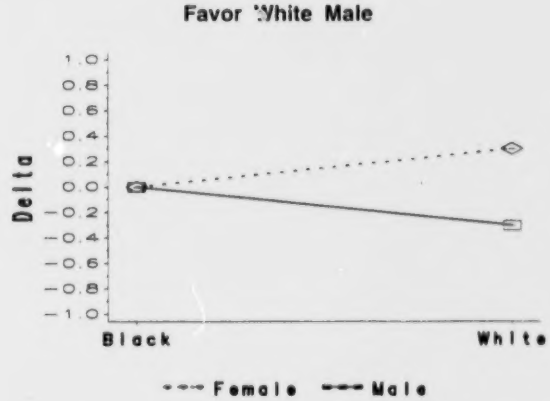
Main Effects



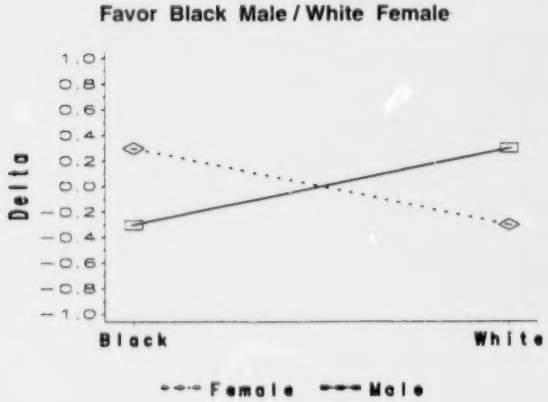


Ordinal Interaction





Disordinal Interaction



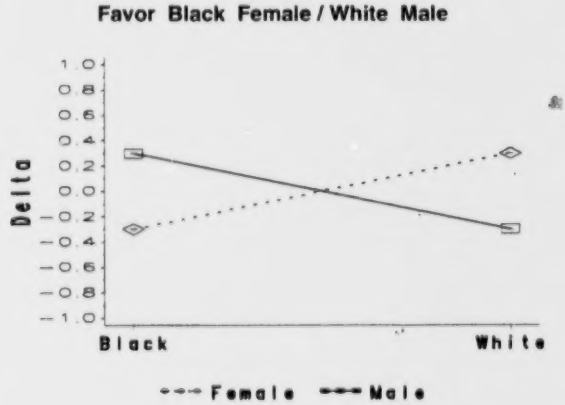


Figure 1. Simulated DIF Items

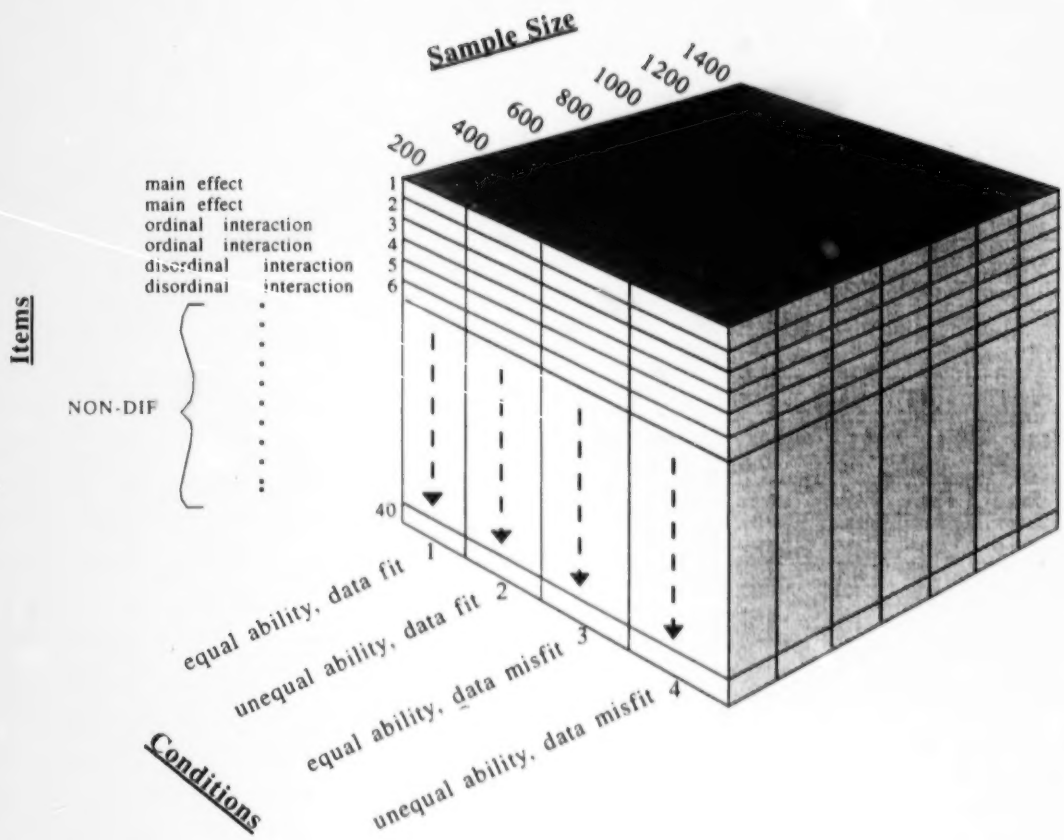


Figure 2. Design of the simulation studies

IV. FINDINGS AND DISCUSSIONS

The findings from the simulation studies are presented and discussed in this section, with a focus on the power and error rate of the method in analyzing interaction-effect DIF. Particular attention will be given to the impact on power and error rate of the interactive relationship between group ability difference, DIF direction (whether the high or low ability group is favored), and model-data misfit.

Detecting ordinal-interaction DIF when data fit the model

Power is defined as the percent of correct rejections or the rejection rate for DIF items. The significance level for rejection is set at .05. Figure 3 displays the rejection rate for items 3 and 4, both with ordinal interaction, when data fit the model. Item 3 favors white females, the low ability group under conditions 2 and 4. Item 4 favors white males, the high ability group under conditions 2 and 4. Each graph has three shaded regions. The bottom region displays the proportion of replications in which both the interaction effect and the main effect (for gender) are significant. The middle region shows the proportion of replications in which only the interaction effect is significant. The top region exhibits the proportion of replications where only the main effect is significant.

The graphs show that, overall, the rejection rate is almost as high for the interaction effect as for the main effect. This is hardly surprising given the pattern of interactions simulated. When the interaction is ordinal, the interpretation of significant main effects may still be permissible in terms of marginal mean differences. "Because the pattern is ordinal, and because parsimonious explanations are valued in science, we would be justified in proceeding to interpret the main effects.... "(Kennedy & Bush, 1978, p.266). Since both the main and interaction effects are interpretable, it may not be inappropriate to combine the two lower regions as the rejection rate for the interaction effect and combine all the three regions as the overall DIF rejection rate for both the main and the interaction effect. And it is not difficult to see that the use of the interaction model not only is more sensitive to the actual DIF pattern, to the extent it captures the interaction effect, but also results in a higher rejection rate than if only a main-effects model is used.

For Item 3, which favors females(the low ability group under unequal ability conditions), the overall rejection rate is higher when there is no group ability

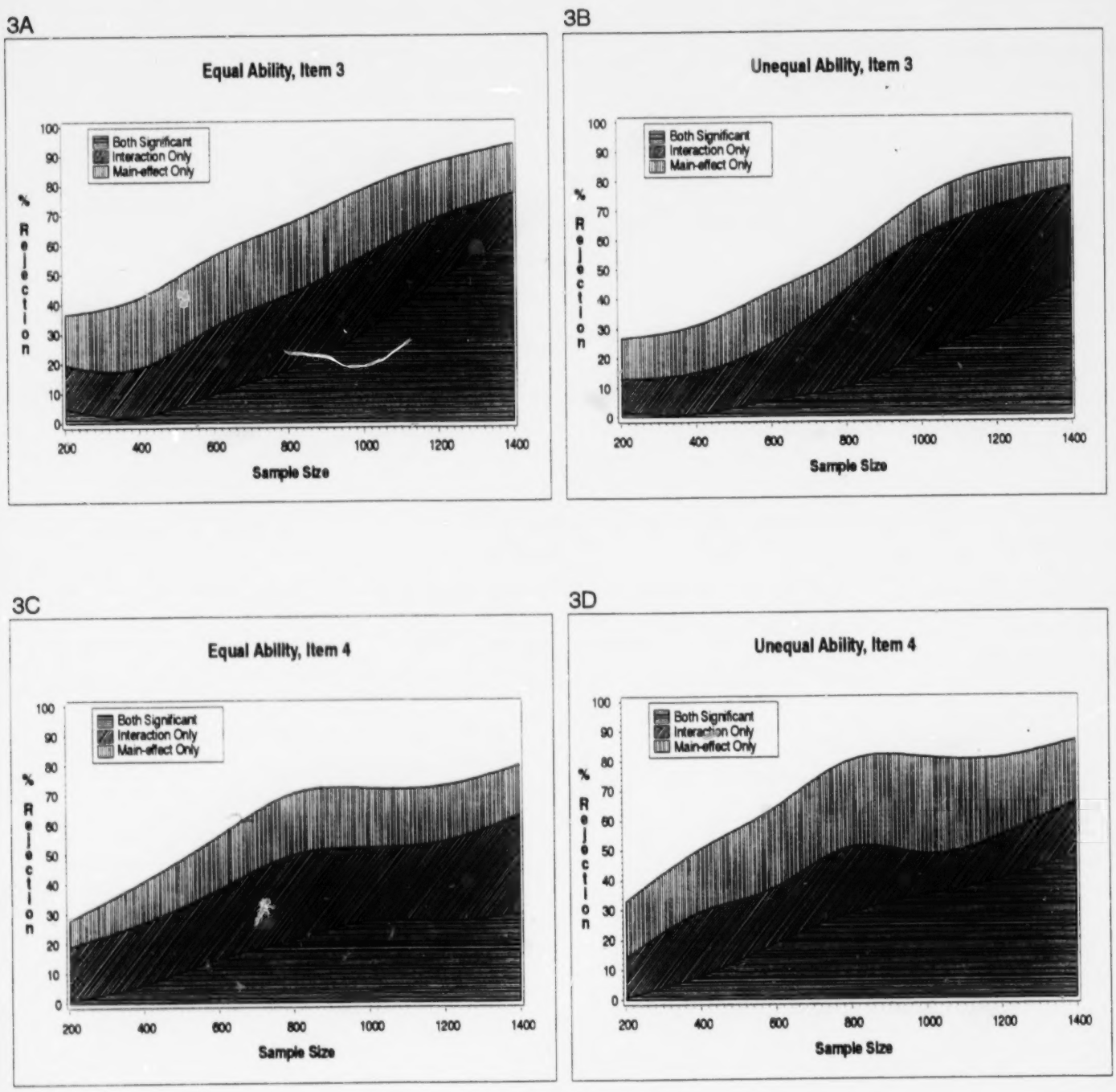


Figure 3. Rejection rate for DIF with ordinal interaction when data fit the model

difference than when there is a group ability difference. For item 4, which favors males (the high ability group under unequal ability conditions), the overall rejection rate is slightly higher when there is a group ability difference than when there is no group ability difference. This suggests that group ability difference interacts with the direction of DIF. Differences in group ability may lead to lower rejection rates when DIF favors the low ability group, and to higher rejection rates when DIF favors the high ability group.

Detecting ordinal-interaction DIF when data do not fit the model

Figure 4 displays the rejection rates for Items 3 and 4 under misfitting conditions. Comparing 4a and 4c with 3a and 3c reveals that the rejection rates are largely the same under equal ability conditions, whether or not data fit the model. Recall that when there is no group ability difference, the impact of the second dimension is evenly distributed among the groups. Hence, it does not have any significant effect on DIF detection.

Comparing 4b and 4d with 3b and 3d, however, we observe considerable differences in rejection rates. The rejection rate for Item 3 is higher, particularly for the main-effect, when data misfit the model than when data fit the model. This is because the item favors the group that the second dimension favors. The advantage of the low ability group on this item is compounded by their advantage on the second dimension. On the other hand, the rejection rate for item 4 is lower, particularly for the main-effect, when data misfit the model than when data fit the model. This could be due to the fact that the item favors the group the second dimension disfavors. The advantage of the high ability group on this item is partially offset by their disadvantage on the secondary dimension.

Figure 5 presents the plot of the means of the group residual means over 100 replications for sample size 1000 for Item 3 and 4 under misfitting conditions. The two graphs on the left (5a and 5c) show that when group ability is equal, the pattern of interaction is much the same as is intended for these two items: No significant difference between males and females among African Americans, but a significant difference among White males and females (see Figure 1). The graphs on the right show that, for Item 3, while the ordinal interaction pattern remains, the main-effect of gender has increased. For Item 4, the pattern of interaction has changed from ordinal to disordinal, and that there appears to be very little difference in marginal means for males and females. In both cases, the changes are due to the differential impact of the secondary dimension on the gender groups.

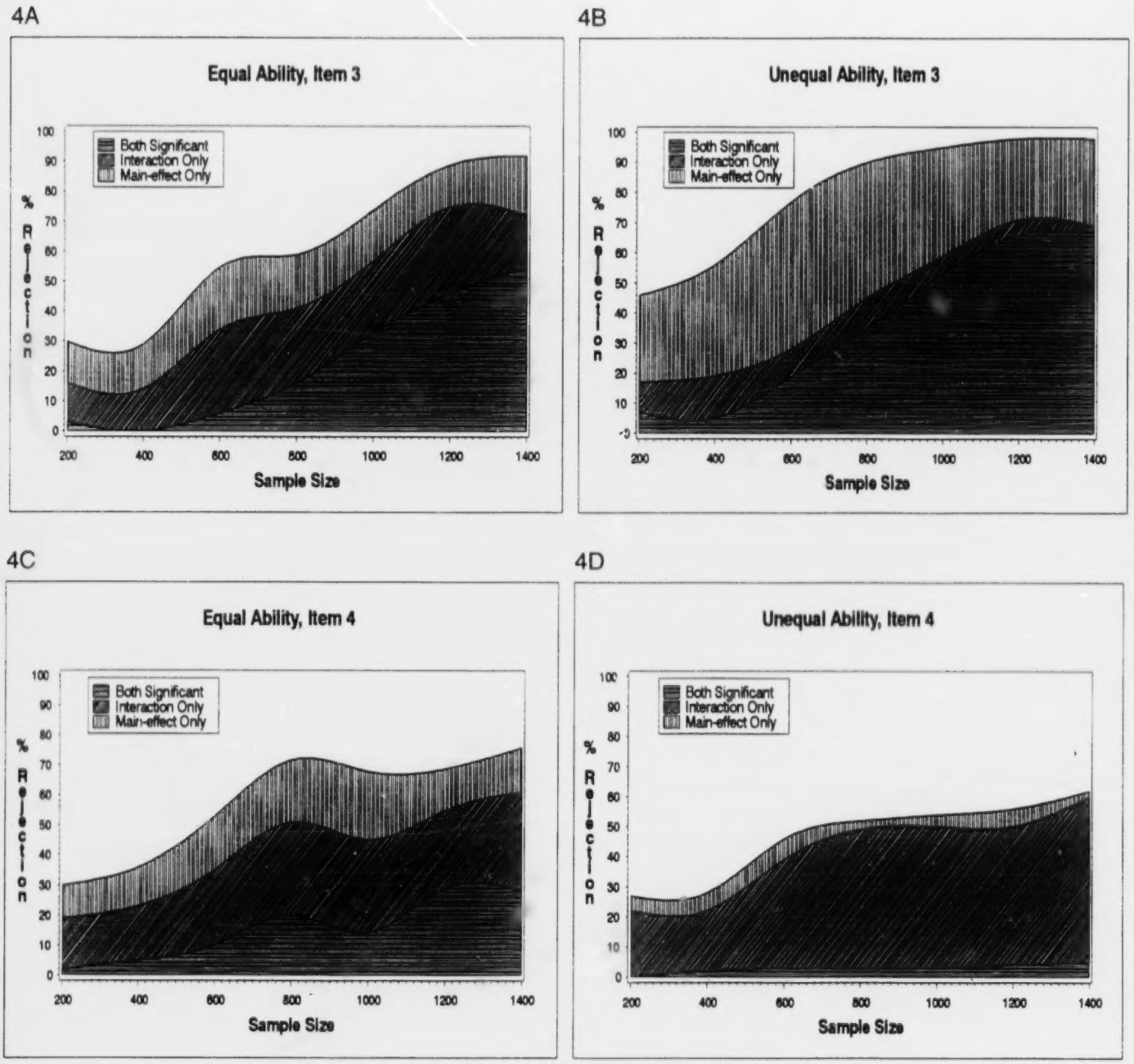


Figure 4. Rejection rate for DIF with ordinal interaction when the model does not fit the data.

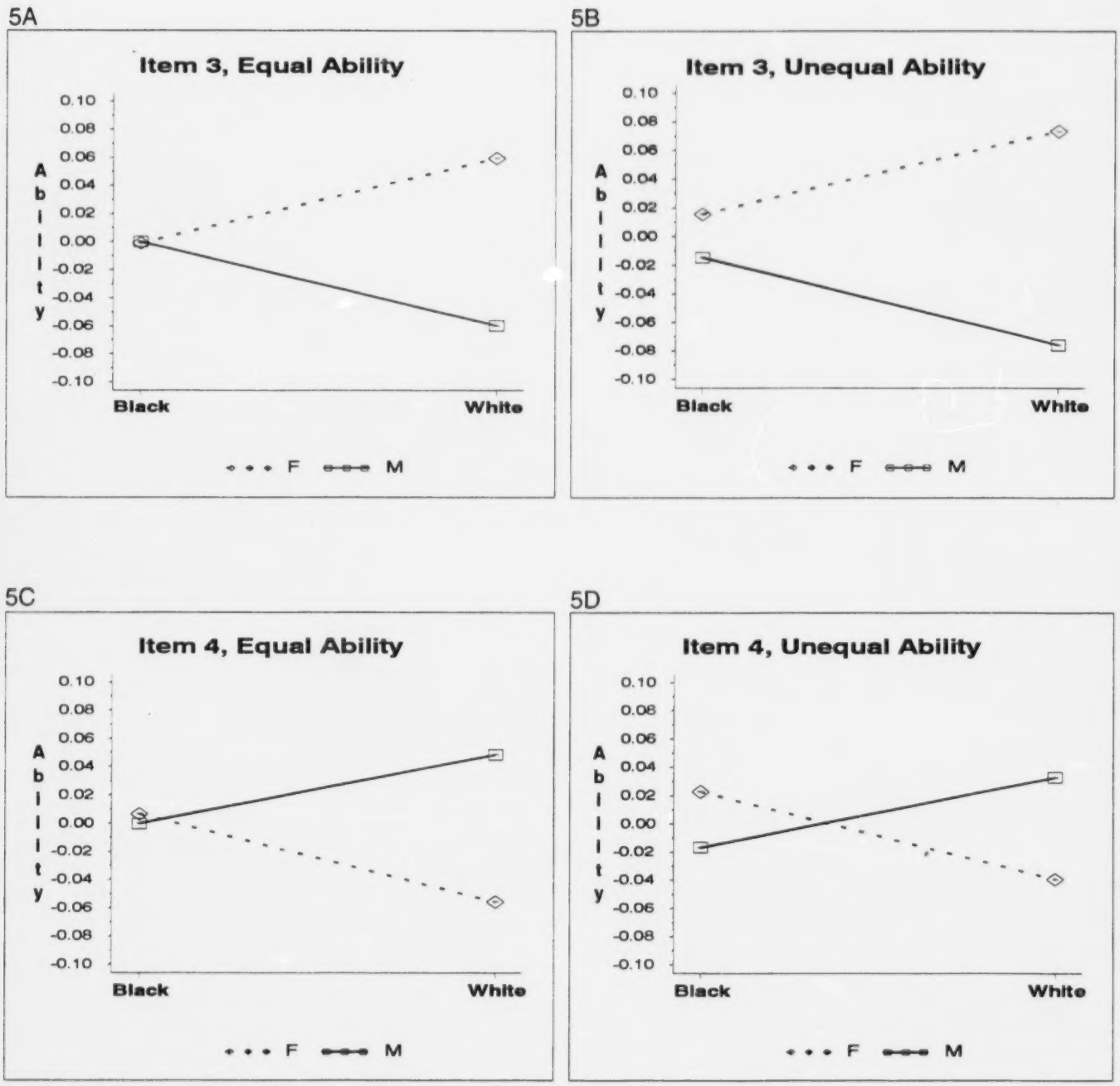


Figure 5. Plot of the means of group residual means over 100 replicationsfor sample size1000 for items 3 and 4 under misfitting conditions.

Detecting DIF with disordinal interaction

Figure 6 presents the rejection rates of the gender by ethnicity interaction effect for Items 5 and 6, both of which were simulated with disordinal interaction DIF. It shows that the rejection rate is almost the same under all four conditions. Neither group ability difference nor model-data misfit seem to have any significant effect on the rejection rate, as the impact of group ability difference and the secondary dimension is evenly distributed among the groups in comparison.

The error rate of the IRT-ANOVA method

Figure 7 presents the plot of the error rate for Item 7, the only non-DIF item simulated as a misfitting item. It also presents the average error rate of all the non-misfitting, non-DIF items (items 8 through 40). The following observations can be made:

1. The error rate is very close to or fluctuates around its nominal level of .05 when there is no group ability difference, whether or not data fit the model (see the two graphs on the left).

2. There is a slight increase in error rate for the gender effect when data fit the model but there is a group ability difference (see the graph on the upper right).

3. There is a considerable increase in error rate for the gender effect when data do not fit the model *and* there is a group ability difference (see graph on the lower right).

4. The error rate for the gender*ethnicity interaction effect and the main effect for ethnicity is close to its nominal level of .05 under all conditions.

It is interesting to note that model-data misfit only affects the error rate for the main-effect of gender because the simulated misfit has a systematic relationship with the levels of this variable. In this case, it is a source of DIF by itself and the "error rate" reflects the extent of DIF caused by departure from unidimensionality. Where such a systematic relationship does not exist, model-data misfit has little or not effect on the error rate.

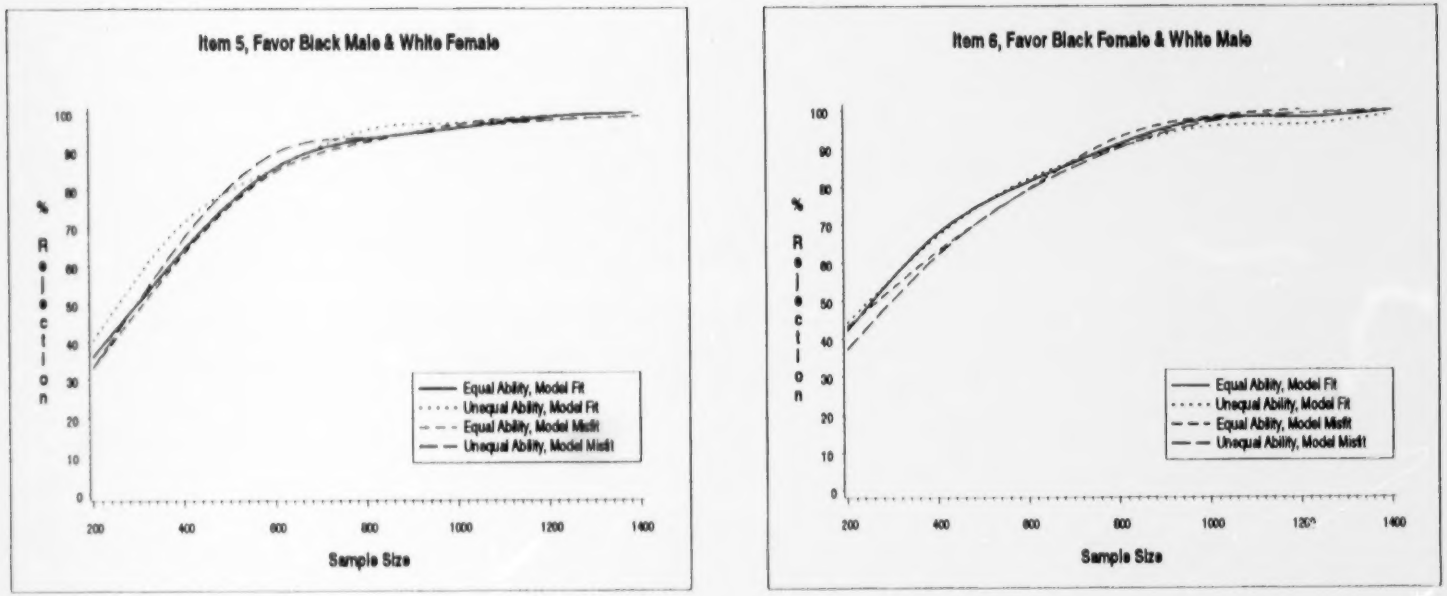
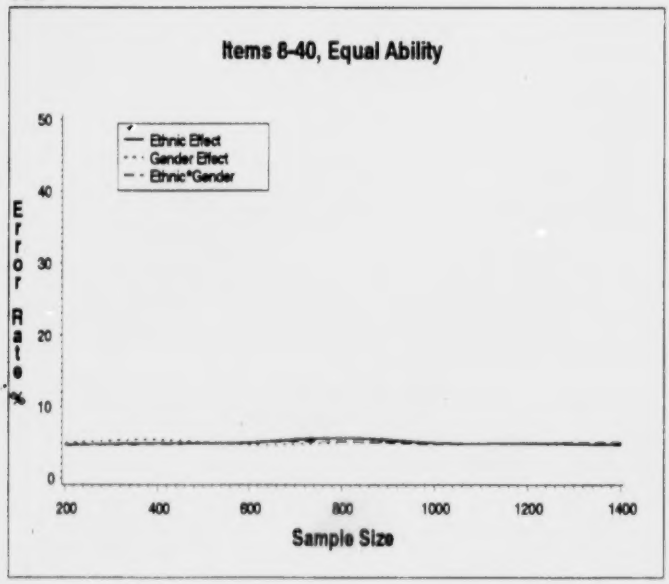

Item 5, Favor Black Male & White Female
% Rejection
Sample Size
Equal Ability, Model Fit
Unequal Ability, Model Fit
Equal Ability, Model Misfit
Unequal Ability, Model Misfit
Item 6, Favor Black Female & White Male
% Rejection
Sample Size
Equal Ability, Model Fit
Unequal Ability, Model Fit
Equal Ability, Model Misfit
Unequal Ability, Model Misfit

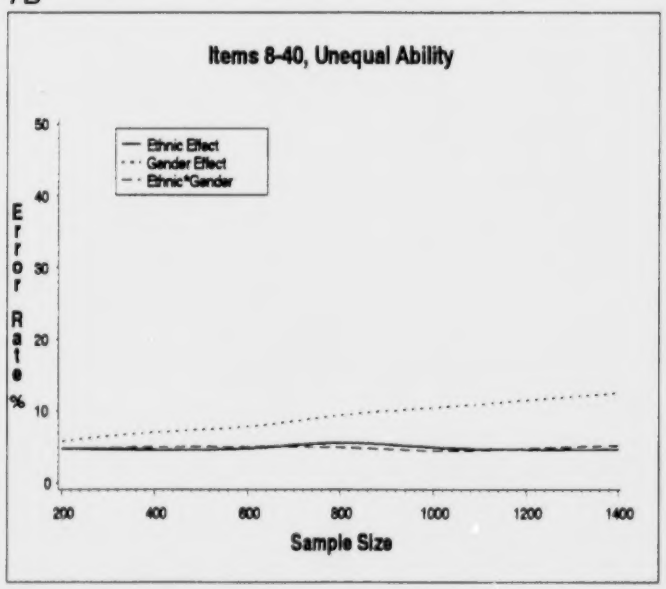
Figure 6. Rejection rate for DIF with disordinal interaction under all conditions.

Data Fit The Model

7A

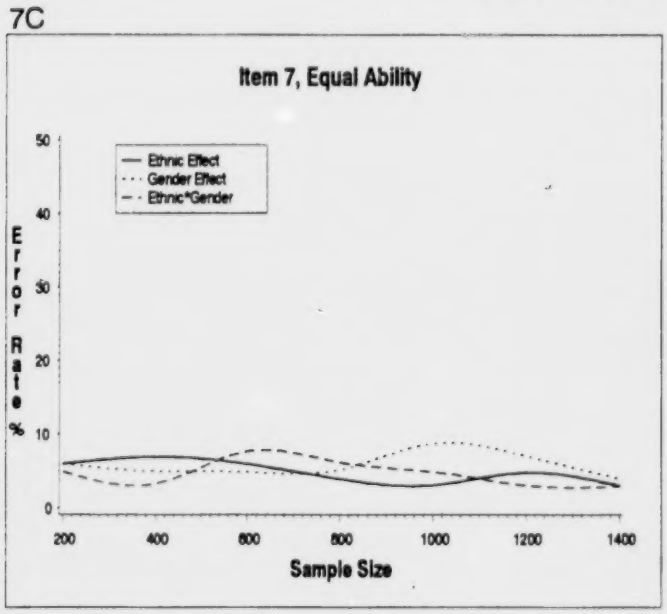


7B



Data Do Not Fit The Model

7C



7D

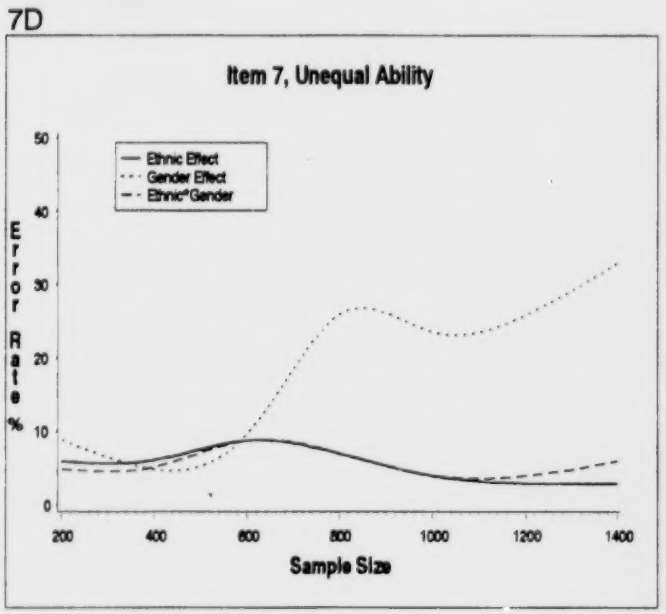


Figure 7. Error rate for misfit and non-misfit items

V. IMPLICATIONS FOR RESEARCH AND APPLICATION

Selection of IRT model. While the procedure makes no assumption about the choice of any particular IRT model or family of models to be used, the selection of an appropriate model is a critical first step in implementing the IRT-ANOVA method. Research is needed to assess the effect of using different IRT models on DIF detection. Of particular interest is the effect of using IRT models that take into account the variation in item discrimination. Item discrimination may vary as a function of other item characteristics, such as item difficulty or small item variance (particularly easy or hard items tend to have low discrimination due to small item variance). But, more often than not, variation in item discrimination results from lack of unidimensionality. It would be interesting to look into the differential consequences, positive or negative, that the use of a two-parameter model might have in each case.

Effect size and statistical significance. It is well known in hypothesis testing situations that statistical significance is not synonymous with practical significance. This is particularly true in large sample situations. One way of evaluating practical significance is to use the observed effect size: the difference in residual means of the groups being compared. It is necessary to interpret the statistical significance values in light of the observed effect size to assess the practical significance of DIF effect.

Raw score or standardized residuals. The use of raw score residuals will enhance the interpretation of the DIF magnitude. The mean residual differences can be interpreted in terms of the raw score scale. For example, if the residual mean (on a 0 to 1 scale) for one group is .1 higher than the residual mean for the other group, it means that the favored group will have 10 percent more examinees responding correctly to the item than the disfavored group due to differential item functioning. The standardized residual may provide better estimates for extremely difficult or easy items. The high residual values that such items are likely to produce are corrected for by their item variances. The disadvantage of using standardized residuals is that the interpretation of the observed effect size is not as straight-forward as the interpretation from the raw score residuals.

Planned comparison. One of the desirable features of the IRT-ANOVA method is the capability to process multiple levels of a factor simultaneously. In situations where several levels of a factor are involved and it may not make sense to make all pairwise comparisons, a planned comparison design may be preferred. For

example, if three ethnic groups (e.g., White, Black, and Hispanic) are involved and only the comparisons between White and Black and White and Hispanic are of interest, it is more efficient (in both statistical and operational sense) to run ANOVA separately for each comparison following concurrent calibration involving all groups.

Unequal group size. In two-group situations, the greater the group size difference given a fixed *n* for the total sample size, the less the statistical power for ANOVA or T-Test (Cohen, 1977). The same does not hold in multiple group situations. Research is underway to investigate the optimal group size difference for DIF detection where multiple groups are involved. In multi-factor situations, unequal group size or cell size may not only affect power, but also necessitate more complicated computational methods to be used to deal with non-orthogonality. It may be desirable to sample the cases to maintain equal or proportional cell sizes (Kirk 1982).

DIF statistics and item fit statistics. When data do not fit the model, it may affect the power or error rate if the misfit has a systematic relationship with the levels of the DIF factor. Misfit caused by departure from unidimensionality could itself be a source of DIF when its impact is not evenly distributed among the groups compared. It is therefore important to interpret DIF statistics in light of item fit statistics and the relationship between the misfit and the DIF factor(s) under investigation.

Testlet and step differential functioning. The procedure can be easily extended to analysis of testlet differential functioning when the items are calibrated at the testlet or cluster level. The procedure may also be adapted to the analysis of possible differential step functioning for items consisting of multiple steps. Residuals can be computed at the score category level by subtracting the category expected score (the probability of scoring in the category) from the category observed score (1 if the examinee scores in the category, 0 otherwise). The residuals can then be plotted for each group and each score category. Potential step differential functioning may be assessed by examining the pattern and the magnitude of the residuals for the groups involved.

VI. CONCLUSION

Item bias, or differential item functioning, has been a perennial concern for test developers and users alike. This is particularly true in high-stake testing situations. The development of theoretically rational and practically feasible methods for DIF analysis has been, and still is, an important area of research. This study explores a new DIF method which is unique to the existing methods in a number of aspects. The most important feature of the method is the capability of simultaneously processing multiple DIF factors, which makes it possible to examine interaction effects and investigate main effects while controlling for possible confounding by other variable(s) included in the design. It is hoped that this study will generate further research interest in the exploration of the simultaneous approach exemplified by the IRT-ANOVA method.

References:

Cole , N. S. & Moss, P. A. (1989). Bias in test use. In Robert L. Linn (Ed): Educational measurement, 3rd edition, pp. 201-220. New York: Macmillan Publishing Company.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Fisher, R. A. (1937). The design of experiments. London: Oliver and Boyd.

Hogg, R. V. & Craig, A. T. (1978). Introduction to mathematical statistics, 4th edition. New York: Macmillan Publishing Co., Inc.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, Winter 1989 (pp. 5-11).

Kennedy, J. J. & Bush. (1978). An introduction to the design and analysis of experiments in behavioral research. University Press of America.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences. Monterey: Brooks/Cole Publishing Company.

Tang, H. (1994). A new IRT-based small sample DIF-method. Paper presented a the Annual Meeting of the Southwest Educational Research Association, January 27-29, 1994, San Antonio, Texas.